



Datawarehousing : fundamentos



Business Intelligence

- “Es un paraguas bajo el que se incluye un conjunto de conceptos y metodologías cuya misión consiste en mejorar el proceso de toma de decisiones en los negocios basándose en hechos y sistemas que trabajan con hechos”

Howard Dresner
(Gartner Group), 1989



B.I.: recursos y herramientas

- Fuentes de datos : warehouses, data marts, etc
- Herramientas de administración de datos
- Herramientas de extracción y consulta
- Herramientas de modelización (Data Mining)



Evolución: Business Data to Business Information

Etapas	Pregunta de Negocio	Tecnología disponible	Proveedores	Características
Data Collection (1960)	¿Cuál fue el total de ventas en Capital Federal y GBA?	Computadoras, cintas, discos	IBM, NCR, etc	Retrospectivo Estático
Data Access (1980)	¿Cuáles fueron las ventas por sucursal en Capital Federal y GBA?	RDBMS SQL	Oracle, Informix, Sybase, etc	Retrospectivo Dinámico



Evolución: Business Data to Business Information

Etapas	Pregunta de Negocio	Tecnología disponible	Proveedores	Características
Data Navigation (1990)	¿Cuál fue el total de ventas en Capital Federal? Drill down a GBA	OLAP DW	Pilot, Discoverer, Arbor, etc	Retrospectivo Dinámico Niveles múltiples
Data Mining (2000)	¿Cómo evolucionarán las ventas en el próximo año?	Algoritmos avanzados Multiprocesadores	Intelligent Miner (IBM) SGI SAS, etc	Prospectivo. Proactivo



Data Warehouse

- Data Warehouse is a subject-oriented, integrated, time-variant, non volatile collection of data in support of management decisions

Bill Inmon (1990)



Subject oriented

- Los datos almacenados en el DataWarehouse proveen información sobre **un tema** en particular en vez de atender la operatoria de gestión de la compañía.



Integrated

- Los datos son volcados al DataWarehouse desde **diferentes fuentes e integrados** en un todo consistente.



Time-variant

- Todos los datos del datawarehouse refieren a un particular momento en el tiempo (como una "foto" o "snapshot").



Non volatile

- Los datos son **estables**. En general siempre se **agregan** datos pero no se quitan . Esto permite análisis **retrospectivos** sobre la marcha del negocio.



Data Warehouse

- "A copy of transaction data specifically structured for query and analysis".

Ralph Kimball



Datos operacionales y Data Warehouse

	Datos operacionales	Data Warehouse
Contenido	Valores elementales	Datos sumariados, derivados
Organización	Por aplicación	Por tema
Estabilidad	Dinámicos	Estáticos hasta su actualización



Datos operacionales y Data Warehouse

	Datos operacionales	Data Warehouse
Estructura	Optimizada para uso transaccional (NORMALIZADA)	Optimizada para queries complejos (DESNORMALIZADA)
Frecuencia de acceso	Alta	Media y baja
Tipo de acceso	Lectura / escritura Actualización campo por campo	Lectura Sumarización



Datos operacionales y Data Warehouse

	Datos operacionales	Data Warehouse
Uso	Predecible Repetitivo	Ad hoc Heurístico
Tiempo de respuesta	Segundos	Segundos a minutos
Cantidad de registros involucrados	A lo sumo decenas	Cientos - miles

La arquitectura de los datos





Problemas con los datos

- Demasiados datos
 - datos corruptos o con ruido
 - datos redundantes (requieren factorización)
 - datos irrelevantes
 - excesiva cantidad de datos



Problemas con los datos

- Pocos datos
 - atributos perdidos (missings)
 - valores perdidos
 - poca cantidad de datos
- Datos fracturados
 - datos incompatibles
 - múltiples fuentes de datos



¿Cuántos datos son necesarios?

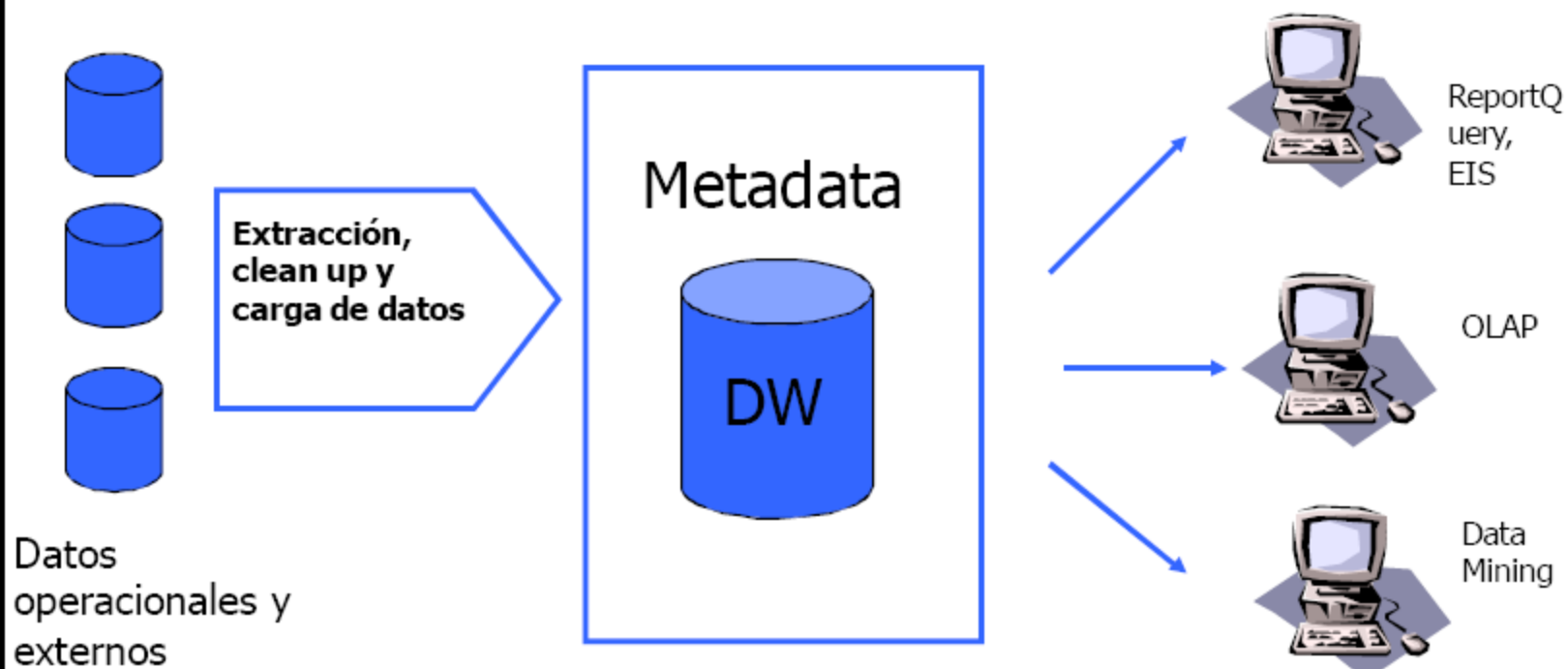
- ¿Cuántas filas?.
- ¿Cuántas columnas?.
- ¿Cuánta historia?
- Regla general : cuanto más datos, mejor
- En la práctica : condicionado a los recursos de obtención y procesamiento.



Data Marts

- Técnicamente es un subconjunto del DW orientado a una finalidad específica de negocio : marketing, finanzas, producción, etc
- El término se utiliza también para identificar soluciones alternativas a un DW corporativo más reducidas y de menor costo y tiempo de implantación.

Explotación del Datawarehouse





Componentes del DW

- Fuentes de datos
- Procedimientos de **E**xtracción
- Procedimientos de **T**ransformación
- Procedimientos de carga (**L**oading)
- Soporte físico de los datos (DBMS)
- Herramientas de explotación : OLAP, reporting, Data Mining, etc.

ETL



Adquisición y limpieza

Objetivos

- Remover datos no necesarios de las fuentes operacionales
- Consolidar representaciones de datos de diferentes fuentes
- Calcular sumalizaciones y variables derivadas
- Resolver problemas de missings y outliers



Metadata

- Provee a los usuarios de información para facilitarles el acceso e interpretación del contenido del DW



Metadata

- Información sobre los datos:
 - Fuentes de datos
 - Descripción de operaciones de transformación
 - Estructura de datos del DW
 - Reglas de clean up
 - Referencias históricas y temporales,etc



La importancia de los metadatos

- Los **metadatos** proveen la vinculación entre los datos y los usuarios de negocio. Describen los datos
- Incluyen los modelos lógicos de datos, el mapeo de los datos a los sistemas transaccionales, el esquema físico de los datos, información de carga, actualización y seguridad, etc.



- Procedimientos (herramientas) destinados a obtener los datos de las fuentes operacionales, limpiarlos, convertirlos a los formatos de utilización y cargarlos en el repositorio final.



Integridad de datos

- Los datos cumplen condiciones de integridad cuando se ajustan a todos los estándares de valor y completitud.
- Todos los datos del DW son correctos
- El DW está completo (no existen más datos fuera de él).



Integridad de datos

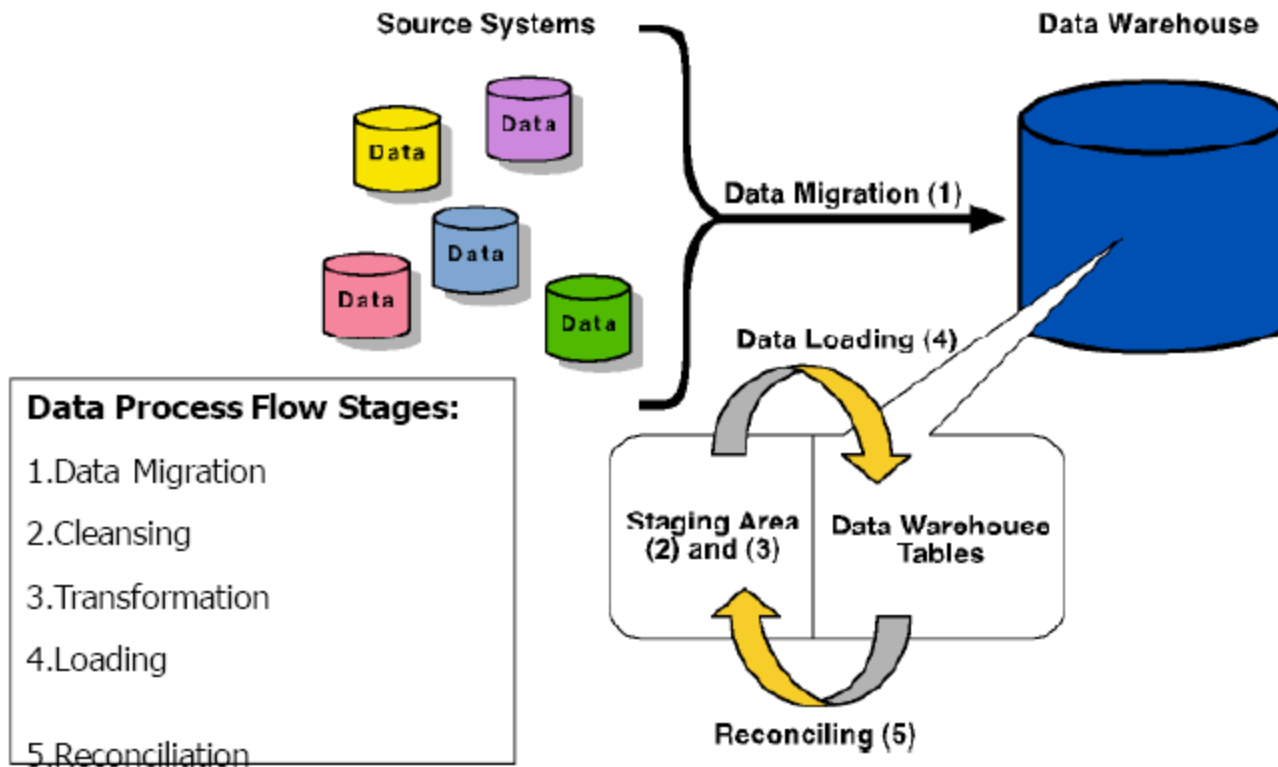
- La credibilidad del DW depende de la integridad de sus datos
- El uso del DW depende de la percepción de los usuarios y de la confianza que tengan en su contenido.
- De la integridad de datos depende el éxito del proyecto.



Controles de Integridad

- Controles de Prevención : controlan la integridad antes de cargar los datos en el DW.
- Controles de Detección : aseguran la exactitud y completitud de la información una vez cargada en el DW.

Data Process Flow





Etapas del proceso ETL

- Migración de datos
- Limpieza
- Transformación (cálculos, agregados, sumarizaciones, de-normalización).
- Carga
- Conciliación - Validación



Migración

- Staging area : área de trabajo fuera del DW.
- El propósito de la migración es mover los datos de los sistemas operacionales a las áreas de trabajo (staging areas).
- NO se debe mover datos innecesarios (control preventivo).



Limpieza (Data cleaning)

- Corregir, estandarizar y completar los datos.
- Identificar datos redundantes
- Identificar valores atípicos (outliers)
- Identificar valores perdidos (missings)



Limpieza (ejemplo)

FIRST_NAME	LAST_NAME	COMPANY_NAME	AREA_CODE	PHONE	STATE
sAM	Adams	boston beer co.	617	3685000	MA
Sam	Adams	Boston beer co.,	617	3685000	MA
Samuel	Adams	Boston Beer Co.	617	3685000	MA
SAMUEL	ADAMS	BOSTON BEER	617	3685000	MA
					
Samuel	Adams	Boston Beer Co.	617	3685000	MA



Limpieza (otros ejemplos)

- Eliminar transacciones con monto = 0 (promociones, regalos)
- Eliminar transacciones anuladas (balance = 0).
- Normalizar nombres de marcas de auto, de direcciones, etc.
- Eliminar fechas de nacimiento inválidas (edad > 100 años o negativa)



Limpieza (actividades)

- Las denominaciones de los sistemas operacionales deben uniformarse y referenciarse con nombres propios de los sistemas de negocios (autodocumentados)
- Cust
- Cust_id
- Cust_nro



Nro de Cliente



Limpieza (actividades)

- Los tipos de dato asociados a cada atributo deben standarizarse y consolidarse para las diferentes fuentes.

- Nombre (A20)
- Nombre (A25)

Nombre
A(25)





Limpieza (actividades)

- Se debe uniformar las tablas de códigos de los sistemas operacionales y simplificar esquemas de codificación
- Datos complejos, que representan varios atributos a la vez, deben ser particionados.



Transformación

- Son procesos destinados a adaptar los datos al modelo lógico del DW
- Se generan “reglas de transformación”.
- Las reglas deben validarse con los usuarios del DW



Transformación

- Generalmente el DW no contiene información de las entidades que - en los sistemas operacionales - son muy dinámicas y sufren frecuentes cambios.
- Si es necesario se utilizan Snapshots (fotos instantáneas)



Transformación

- La des-normalización de los datos tiene como propósito mejorar la performance.
- Otro propósito es el de reflejar relaciones estáticas, es decir, que no cambian en una perspectiva histórica. Por ejemplo: producto - precio vigente al momento de facturación.



Transformación (sumarizaciones)

- Los datos sumarizados aceleran los tiempos de análisis.
- Las sumarizaciones también ocultan complejidad de los datos.
- Las sumarizaciones pueden incluir joins de múltiples tablas
- Las sumarizaciones proveen múltiples vistas del mismo conjunto de datos detallados (dimensiones).



Sumarizaciones (mantenimiento)

- El mantenimiento de las sumarizaciones es una tarea crítica.
- El DW debe actualizarlas a medida que se cargan nuevos datos.
- Debe existir alguna forma de navegar los datos hasta el nivel de detalle (drill down).
- La definición de la granularidad es un problema serio de diseño.



El nivel de granularidad: problema de diseño del DW

- ¿Cuál es la unidad de tratamiento (fila)
- ¿Qué es un cliente? Una cuenta, un individuo, una familia
- ¿Cómo se sumariza la dimensión tiempo? Días, semanas, meses ...?



Carga (Loading)

- Dos aproximaciones:
 - Full Refresh
 - Incremental
- Aunque el Full Refresh parece más sólido desde el punto de vista de la integridad de los datos, a medida que crece el DW se vuelve cada vez más difícil de realizar.



Controles de detección

- La validación de la carga del DW identifica problemas en los datos no detectados en las etapas anteriores.
- Existen dos maneras de hacer la validación:
 - completa (al final del proceso)
 - por etapas a medida que se cargan los datos



Controles de detección

- Los controles incluyen reportes que comparan los datos del DW con las fuentes operacionales a través de:
 - totales de control
 - número de registros cargados
 - valores originales vs valores limpios (transformados), etc.



Herramientas ETL

- Pueden ser procesos manuales diseñados a medida (queries SQL, programas en Visual Basic, etc).
- Existen herramientas que proporcionan interfaces visuales para definir joins, transformaciones, agregados, etc. sobre las plataformas mas comunes.



Modelado de datos



La pregunta central

¿De qué modo deben diseñarse las bases de datos que conforman un Data Warehouse para soportar eficientemente los requerimientos de los usuarios?



¿Por qué es importante?

- Visualización del universo del negocio
- Modelo de abstracción de las “preguntas” que los usuarios necesitan responder
- Diseño del plan de implantación del Data Warehouse



Dos técnicas

Modelo E-R

- Entidades
- Atributos
- Relaciones

Modelo dimensional

- Hechos
- Dimensiones
- Medidas

Modelo E-R

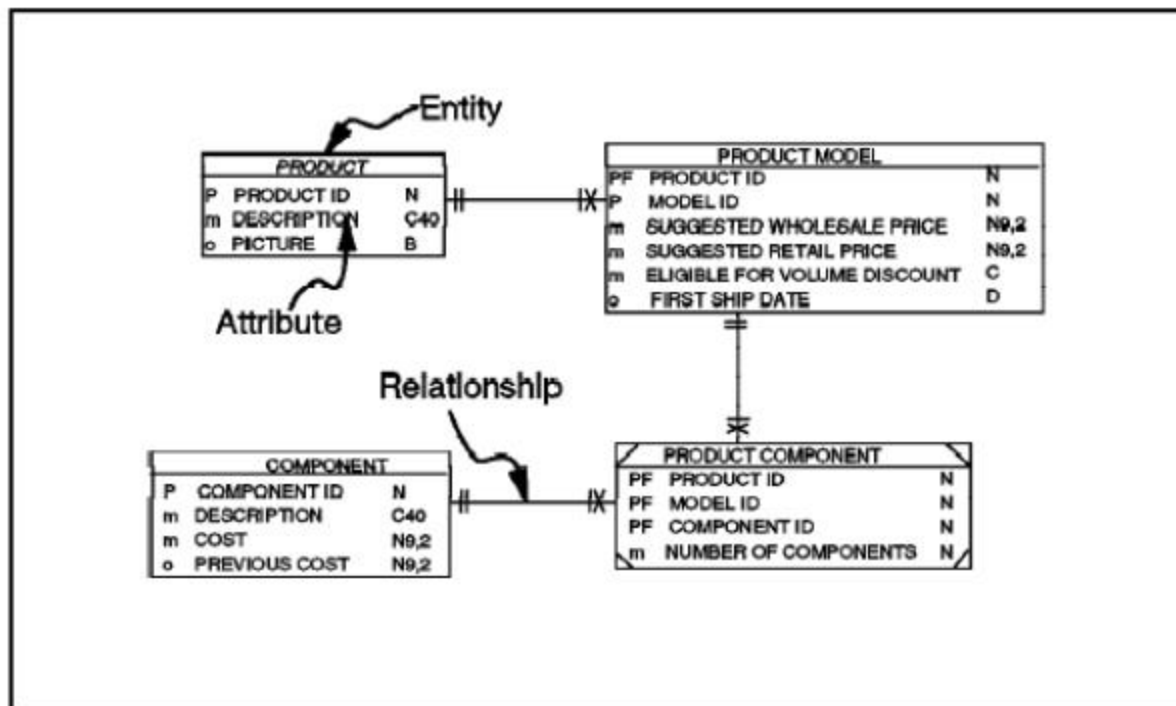


Figure 12. A Sample ER Model. Entity, relationship, and attributes in an ER diagram.



Modelo dimensional: HECHOS

- Hechos : colección de items de datos y datos de contexto. Cada hecho representa un item de negocio, una transacción o un evento
- Los hechos se registran en las tablas CENTRALES del DW



Modelo dimensional: DIMENSION

- Una dimensión es una colección de miembros o unidades o individuos del mismo tipo
- Cada punto de entrada de la tabla de HECHOS está conectado a una DIMENSION
- Determinan el contexto de los HECHOS



Modelo dimensional: DIMENSIONES

- Se utilizan como parámetros para los análisis OLAP
- Dimensiones habituales son:
 - Tiempo
 - Geografía
 - Cliente
 - Vendedor



Modelo dimensional: DIMENSIONES - Miembros

Dimensión	Miembro
Tiempo	Meses, Trimestre, Años
Geografía	País, Región, Ciudad
Cliente	Id Cliente
Vendedor	Id Vendedor

Modelo dimensional

DIMENSIONES - Jerarquía

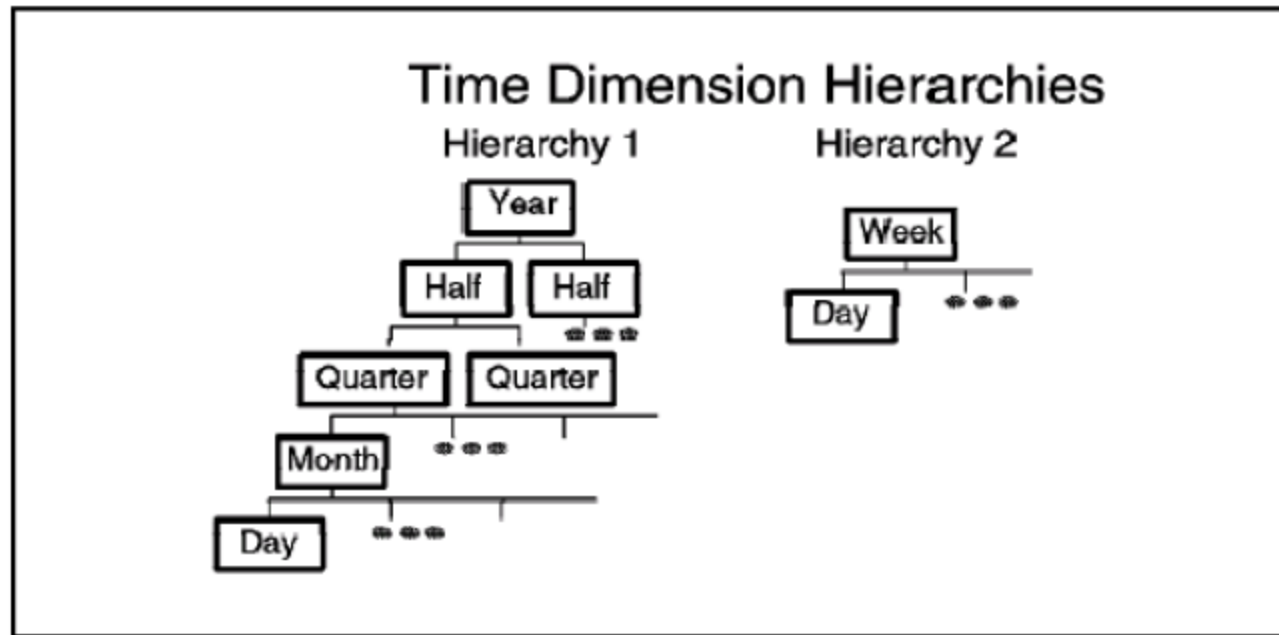


Figure 14. Multiple Hierarchies in a Time Dimension



Modelo dimensional

DIMENSIONES : Medidas

- Medida : es un atributo numérico de un hecho que representa la performance o comportamiento del negocio relativo a la dimensión
- Ejemplos:
 - Ventas en \$\$
 - Cantidad de productos
 - Total de transacciones, etc.

Visualización de un modelo dimensional

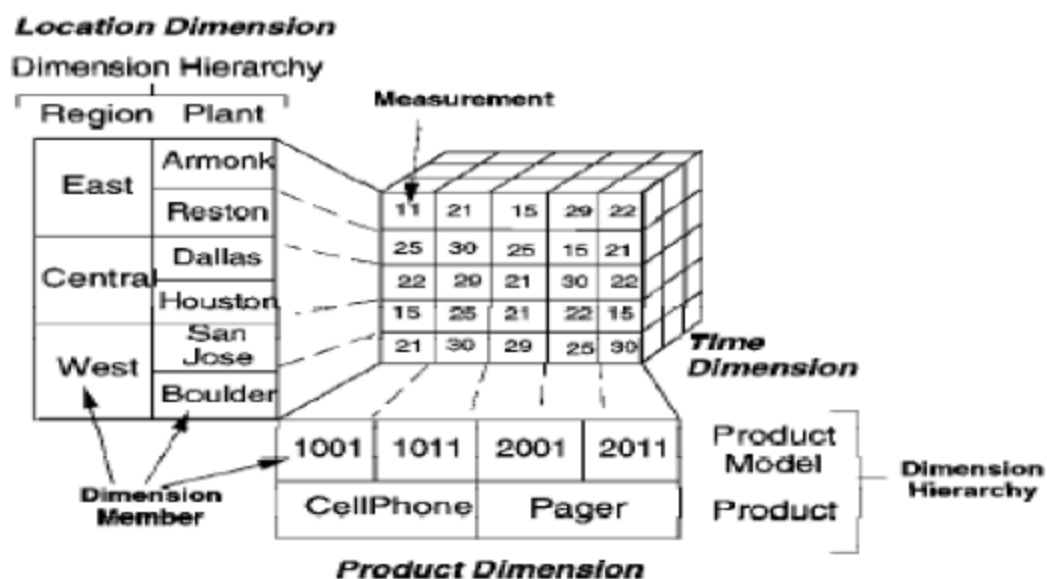


Figure 15. The Cube: A Metaphor for a Dimensional Model



DW - OLAP

El modelo dimensional es ideal para soportar las 4 operaciones básicas de la tecnología OLAP:

- Relacionadas con la granularidad: ROLL UP - DRILL DOWN
- Navegación por las dimensiones : SLICE - DICE



Example: Roll Up and Drill Down

\$ of Anheuser-Busch by drinker/bar

	Jim	Bob	Mary
Joe's Bar	45	33	30
Nut-House	50	36	42
Blue Chalk	38	31	40



Roll up
by Bar

\$ of A-B / drinker

Jim	Bob	Mary
133	100	112



Drill down
by Beer

\$ of A-B Beers / drinker

	Jim	Bob	Mary
Bud	40	29	40
M'lob	45	31	37
Bud Light	48	40	35

Drill Down - Roll Up

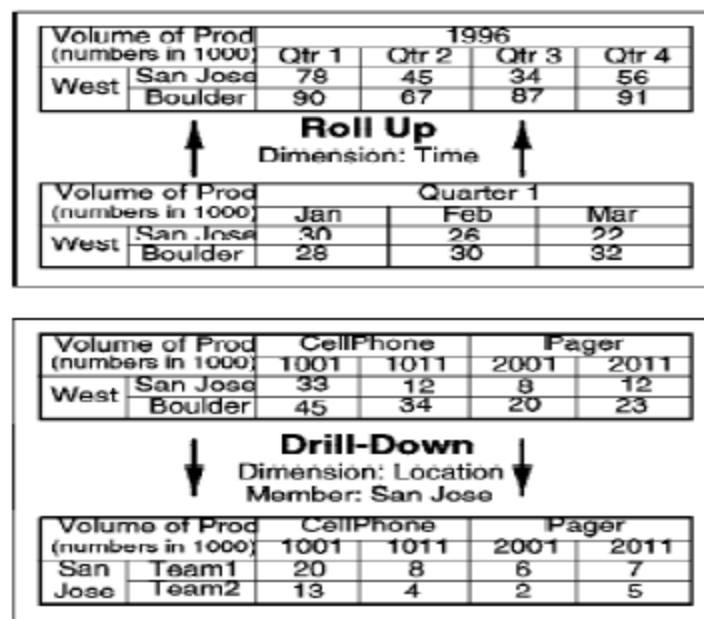


Figure 16. Example of Drill Down and Roll Up



Slice and Dice

- Slice: es un subconjunto del "array multidimensional" que tiene un único valor para una o más dimensiones. Es una "rebanada" del cubo
- Dice: es como el "slice" pero para 2 ó más valores de una o más dimensiones



Modelos básicos dimensionales

STAR



SNOWFLAKE





Star

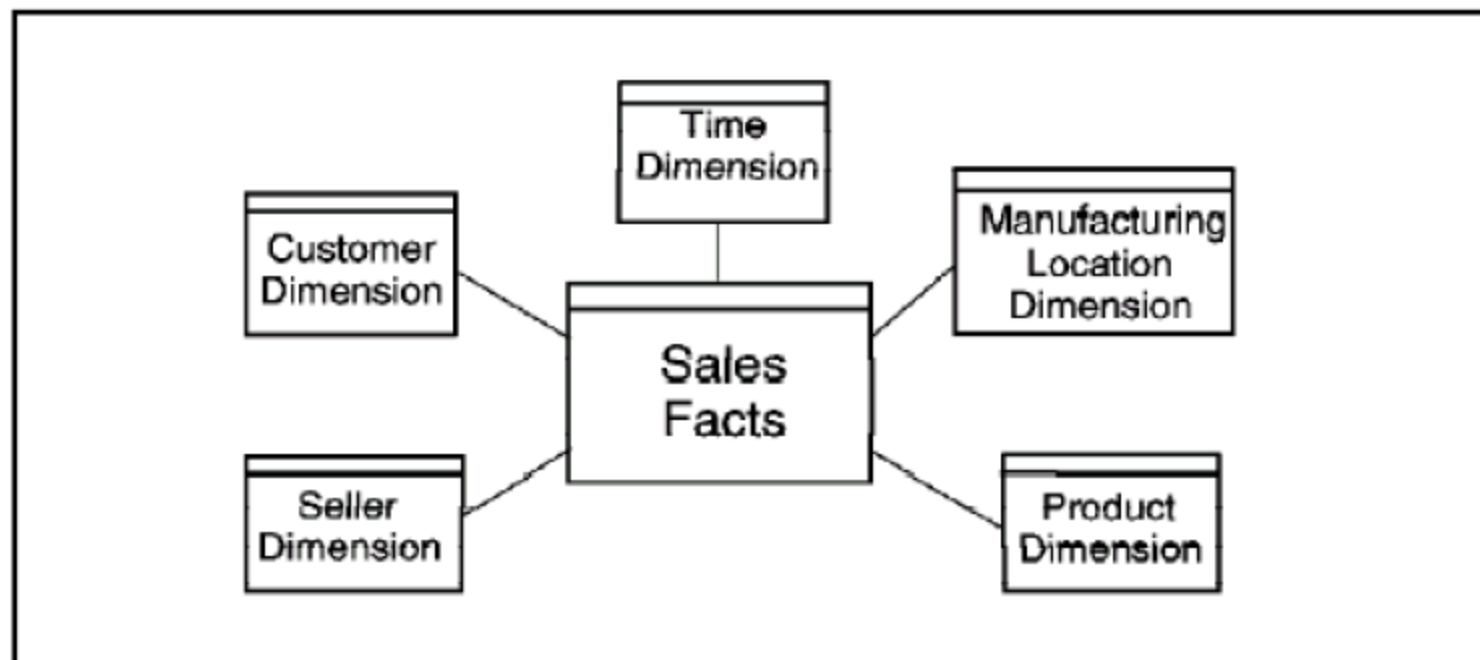
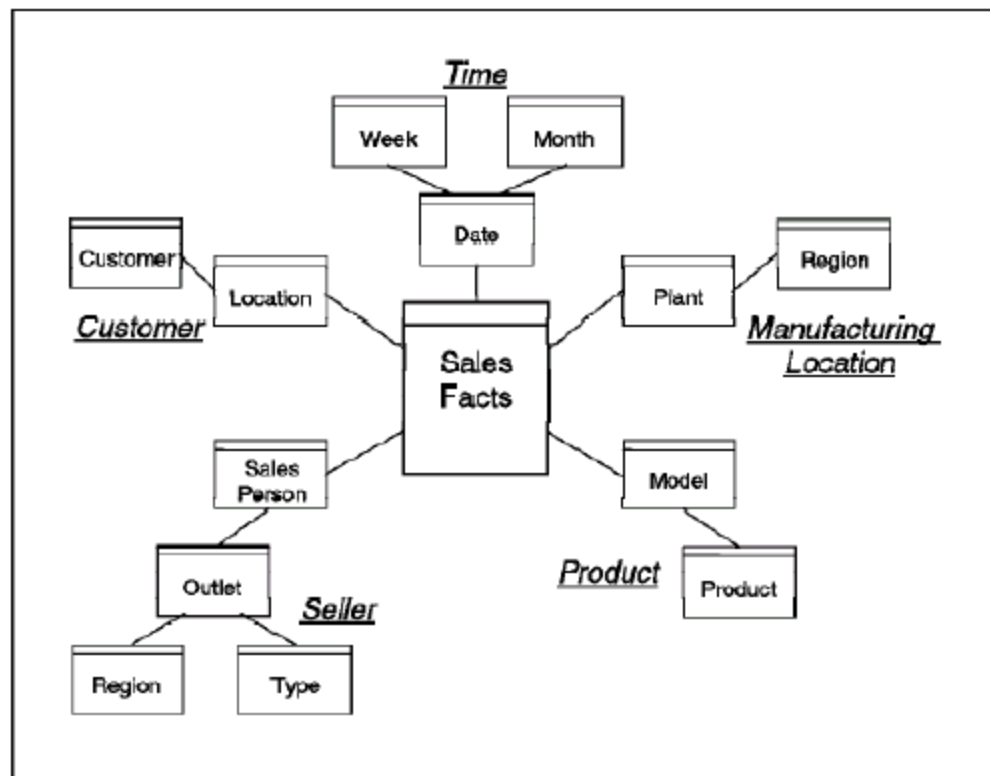




Figure 18. Star Model.





E-R - Modelo dimensional

- El modelo dimensional puede verse como un caso particular del modelo de ER
- Foreign keys  Dimension
- Hecho  Entidad



Presentación

- Esta presentación fue armada utilizando, además de material propio, material contenido en los manuales de Oracle y material provisto por los siguientes autores
- Siblberschat, Korth, Sudarshan - Database Systems Concepts, 6th Ed., Mc Graw Hill, 2010
- García Molina/Ullman/Widom - Database Systems: The Complete Book, 2nd Ed., Prentice Hall, 2009
- Elmasri/Navathe - Fundamentals of Database Systems, 6th Ed., Addison Wesley, 2011
- Ing. Maria del Rosario Bruera y Lic. Néstor Martínez